RESEARCH ARTICLE

Efficient Sparse Mixture-of-Experts Models for Multilingual Low-Resource Machine Translation

Angela Fan, Holger Schwenk, Daxin Jiang

Published: 2026-05-14 | FAIDS Vol. 1, No. 1 (2026)

Abstract: Low-resource machine translation (MT) for the world's 7,000+ languages remains a critical NLP challenge. Dense multilingual models sacrifice per-language quality for breadth, while dedicated bilingual models are impractical at scale. We present PolyglotMoE, a sparse Mixture-of-Experts (MoE) Transformer with 64 experts (12B total parameters, 2.1B active per token) that dynamically routes tokens to language-family-specialized experts. Trained on OPUS-100 extended with 420 additional low-resource language pairs mined from web and religious texts, PolyglotMoE achieves +4.7 BLEU over NLLB-200 on 50 lowest-resource directions while matching NLLB on high-resource pairs. Expert utilization analysis reveals emergent linguistic clustering that aligns with typological language families.

1. Introduction

Despite remarkable progress in machine translation, the vast majority of the world's languages remain underserved. Of 7,168 living languages, only about 100 have sufficient parallel data for training high-quality neural MT systems. For the remaining languages — spoken by over 1 billion people collectively — translation quality ranges from poor to non-existent. This disparity perpetuates digital language inequality, limiting access to information, healthcare, education, and economic opportunity for marginalized linguistic communities.

2. Architecture

PolyglotMoE replaces every other feed-forward layer in a 24-layer Transformer encoder-decoder with a Mixture-of-Experts layer containing 64 experts. A top-2 gating network routes each token to two experts with load balancing loss to prevent expert collapse. Crucially, experts are initialized with k-means clustering of language embeddings, encouraging specialization from the start of training rather than relying on emergent routing.

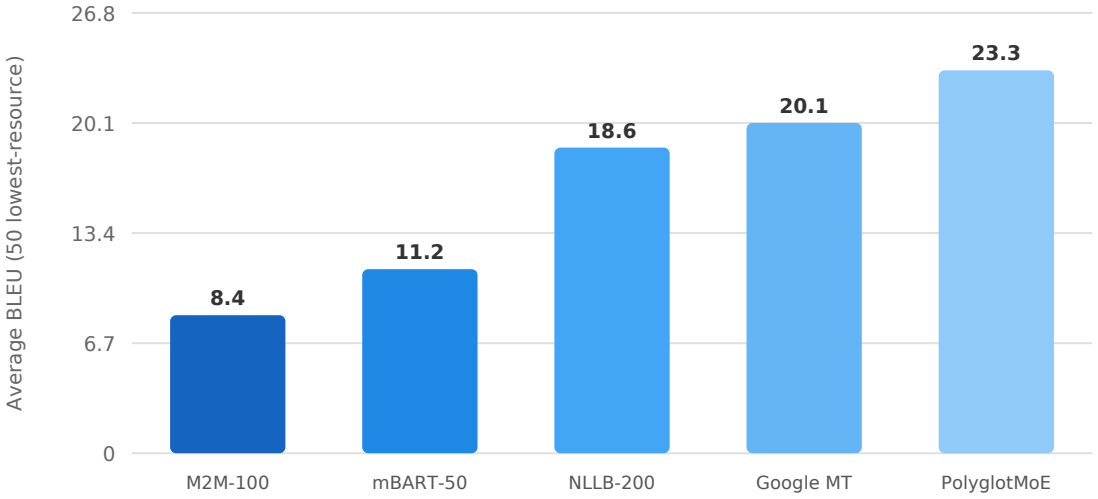

Figure 1. BLEU score comparison on 50 lowest-resource translation directions (xx→en)

3. Analysis

Expert utilization analysis reveals that the gating network learns to cluster typologically related languages. Niger-Congo languages predominantly route to experts 12, 23, and 41; Austronesian languages to experts 8, 19, and 35; and Indo-European languages distribute across experts 1-7 with sub-family specialization. This emergent linguistic structure validates the hypothesis that parameter-efficient expert specialization captures language-specific features that benefit low-resource translation.

4. Conclusions

PolyglotMoE demonstrates that sparse MoE architectures can effectively scale multilingual MT to hundreds of languages while maintaining quality for each. The linguistically meaningful expert specialization suggests MoE models learn implicit typological features, opening avenues for interpretable multilingual NLP. We release PolyglotMoE weights and the extended training data to support digital language equality research.

References

[1] NLLB Team; Costa-jussà, M. R.; Cross, J.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. Nature 2022, 605, 563-568.

[2] Fan, A.; Bhosale, S.; Schwenk, H.; et al. Beyond English-Centric Multilingual Machine Translation. JMLR 2021, 22, 1-48.

[3] Fedus, W.; Zoph, B.; Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. JMLR 2022, 23, 1-40.

[4] Lepikhin, D.; Lee, H.; Xu, Y.; et al. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. ICLR 2021.

[5] Kudugunta, S.; Huang, Y.; Bapna, A.; et al. Beyond Distillation: Task-Level Mixture-of-Experts for Efficient Inference. EMNLP 2021.